Statystyczna Analiza Danych – laboratorium

Analiza składowych głównych (PCA)

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 13
2/3 czerwca 2022

Idea zajęć – co i po co będziemy robić?

- Metody czynnikowe stanowią zbiór metod i procedur statystycznych pozwalających na redukcję dużej liczby zmiennych do kilku wzajemnie nieskorelowanych czynników.
- Za ich pomocą można zachować stosunkowo duża część informacji zawartych w zmiennych pierwotnych.
- Jednocześnie każda z tych metod niesie inne treści merytoryczne. W trakcie laboratorium zaznajomimy się z analizą składowych głównych (PCA).

Dwa modele metod czynnikowych

- Model **klasyczny**, w którym wariancję całkowitą zmiennych dzieli się na wariancję wspólną i wariancję specyficzną (klasyczna analiza czynnikowa FA – tym nie zajmujemy się podczas SAD)
- Model **komponentowy**, w którym nie uwzględnia się struktury wariancji (metoda składowych głównych PCA).
- Żadna z tych metod nie jest zasadna, jeśli zmienne nie są ze sobą skorelowane!

Cele

- **Redukcja liczby zmiennych** bez istotnej straty zawartych w nich informacji
- **Transformacja** układu zmiennych w nowy układ czynników głównych
- **Ustalanie wag** określających znaczenie, jakie należy przypisać poszczególnym zmiennym podczas analizy
- **Ortogonalizacja przestrzeni**, w której rozpatrywane są obiekty będące przedmiotem analizy
- **Opis zjawisk** za pomocą nowych kategorii zdefiniowanych przez czynniki, **tworzenie skal i miar** z kilku zmiennych

Przykłady zastosowań

- Kiedy interesuje nas **eksploracja** i rozpoznanie struktury zbioru danych.
- Gdy nie posiadamy modelu „głębokiej" **struktury czynników** wyjaśniających związki między danymi.
- Gdy potrzebujemy **zredukować zbiór zmiennych** skorelowanych ze sobą do wykorzystania ich w postaci zagregowanej w późniejszych etapach analizy.
- Gdy chcemy stworzyć **skalę, indeks, miernik** ukrytego zjawiska i jednoznacznie wyliczyć jego wartość.

Przykładowe pytania i zagadnienia badawcze

- Stworzenie indeksu kapitału społecznego (FA)
- Wypowiedzenie się na temat postawy respondentów w oparciu o wiele stwierdzeń dotyczących jednego zagadnienia (np. zadowolenia ze spędzania czasu wolnego) (FA lub PCA)
- Stworzenie agregatowej zmiennej z wartości pomiarów potrzebnej do dalszej analizy (PCA)
- Stworzenie zmiennej opisującej objawy depresji, do wykorzystania w regresji liniowej, celem uniknięcia silnego skorelowania zmiennych (PCA)

Analiza składowych głównych (PCA)

- Stanowi metodę transformacji zmiennych pierwotnych we wzajemnie ortogonalne nowe zmienne, tzw. składowe główne
- Służy redukcji wymiaru przestrzeni cech oraz pogrupowaniu ich w podzbiory
- Dzięki niej można graficznie zaprezentować konfigurację porównywanych zmiennych

PCA – ogólna charakterystyka

- Zmienne pierwotne poddaje się standaryzacji, więc ich wariancje są sobie równe
- Nowa agregatowa zmienna powinna wyjaśniać maksymalną ilość wariancji zmiennych pierwotnych
- Wariancja nowej zmiennej agregatowej jest nazywana wartością własną (*eigenvalue*)
- Zbiór danych powinien być jednorodny (brak obserwacji odstających)

Zapis formalny modelu

$$PC_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{ik}X_k$$

$$\sum_{j=1}^{k} w_{ij}^2 = 1$$

- Współczynniki w przy zmiennych X stanowią wagi, jakie przypisuje się zmiennym w tworzeniu głównej składowej
- Zakładamy, że poszukiwane czynniki są niezależne i mają wystandaryzowany rozkład normalny

Wyznaczanie współczynników

- Wartości wektora w są tak dobierane, żeby maksymalizować wariancję PC
- Szukamy wartości własnych następującego równania:

$$|R - \lambda I| = 0$$

- R – macierz korelacji k zmiennych wyjściowych
- Λ – wektor zawierający wartości własne o wymiarach kxk
- Wariancją i-tej składowej jest i-ta wartość własna

Wyznaczanie współczynników – cd

- Każdej wartości własnej możemy przypisać wektor własny macierzy o postaci:

$$Rw_i = \lambda_i w_i$$

- w_i – wektor własny macierzy korelacji
- Wartości składowe tego wektora stanowią wartości współczynników stojących przy zmiennych pierwotnych; ich kombinacja tworzy nowe zmienne: składowe główne
- Pułapka: utworzona kombinacja liniowa jest zależna od jednostek miary i rzędów wielkości poszczególnych zmiennych (należy standaryzować zmienne!)

Ogólne zasady wyboru liczby składowych

- Dążymy do odtworzenia maksymalnej ilości informacji z pierwotnego zbioru zmiennych
- W praktyce wybieramy liczbę składowych, które łącznie wyjaśniają powyżej 70% zmienności zmiennych pierwotnych
- Nie uwzględniamy tych składowych, dla których wartości własne są niższe od średniej
- Można opuścić główne składowe, dla których wartości własne są niższe od 1 (symulacje wskazują, że lepszym progiem jest 0,7)
- Opuszczamy składowe, które mają mniejszy udział w wariancji niż 5%

Metody wyboru liczby składowych

Do wyboru optymalnej liczby składowych można stosować następujące metody:

- Metodę procentu wariancji tłumaczonej przez czynniki
- Metodę wartości własnych większych od jedności
- Metodę wykresu osypiska

Ale i tak ostateczna decyzja jest subiektywnym wyborem badacza

Metoda wartości własnej większej od jedności

- Jest to najczęściej spotykana metoda: każda składowa powinna wyjaśniać zmienność co najmniej jednej zmiennej pierwotnej
- Polecana, jeśli liczba zmiennych jest większa niż 20
- W przypadku analiz na mniejszych zbiorach danych, metoda ta ma tendencję do wybierania zbyt małej liczby składowych

Metoda procentu wariancji tłumaczonej

- Liczbę wybranych składowych ustala się na podstawie procentu wariancji przez nie tłumaczonej
- Dążymy do odtworzenia co najmniej 70% wariancji (niższe wartości w przypadku dużych zbiorów danych)
- Żadna następna składowa poza wybranymi przez nas nie tłumaczy więcej niż 5% wariancji.

Metoda osypiska

- Najpierw sporządzamy wykres, na którym na osi poziomej umieszczamy kolejne składowe, natomiast na osi pionowej ich wartości własne
- Szukamy punktów załamania, w których zmienia się kąt załamania krzywej (zaczynają się kolejne rumowiska)
- Miejsce punktu załamania określa maksymalną liczbę składowych kwalifikujących się do dalszej analizy
- Metoda ta pozwala włączyć do analizy większą liczbę składowych niż metoda wartości własnych większych od 1

Nazwy składowych głównych

- Jeśli składowe główne mają być użyte np. w regresji liniowej (i następnie użyte do interpretacji), dobrze jest nadać im nazwy
- Dla każdej składowej wybieramy kilka zmiennych o najwyższych ładunkach
- Następnie próbujemy nadać wspólną nazwę w oparciu o te zmienne danej składowej

Rotacja czynników

- Uzyskana macierz ładunków czynnikowych często nie jest jedynym możliwym rozwiązaniem analizy czynnikowej
- Można wygenerować nieskończenie wiele różnych macierzy ładunków poprzez obrót układu wzajemnie ortogonalnych osi
- Rotacja ma pomóc w znalezieniu układu, który będzie prostszy w interpretacji
- Istnieją dwie grupy metod rotacji: ortogonalne i ukośne

Rotacje ortogonalne

- Polegają na znalezieniu ortogonalnej macierzy transformacji
- Najbardziej znane metody to **varimax** i **quartimax**
- Varimax minimalizuje liczbę zmiennych potrzebnych do wyjaśnienia danego czynnika
- Quartimax minimalizuje liczbę czynników potrzebnych do wyjaśnienia danej zmiennej

Rotacje ukośne

- Macierz ładunków staje się macierzą wzorców zachowań
- Do wyznaczenia korelacji czynników wykorzystuje się wagi nadane poszczególnym czynnikom F (**promax** – rotacja skośna)

Stosowalność

- Pierwsza główna składowa wyjaśnia najwięcej zmienności wyjściowego zbioru danych jednak nie zawsze jest głównym celem zainteresowań analityków
- Np. W badaniach psychiatrycznych pierwsza główna składowa dostarcza informacji o ostrożności (intensywności) objawów (choroby, anomalii itp.) a kolejne świadczą o wzorcu tych objawów

PCA do wykrywania obserwcji odstających

Ostatnie główne składowe można wykorzystać dla znalezienia obserwacji nietypowych, odstających:

- Przedstawiając ostatnie dwie główne składowe na wykresie można zidentyfikować obserwacje leżące z dala od innych
- Takie obserwacje są podejrzane o bycie obserwacjami odstającymi, ponieważ to one dodają dodatkowy wymiar do głównych składowych
- Podobnie jeśli przedstawimy główną składową na wykresie (histogram), to występowanie bardzo małych wartości lub bardzo dużych będzie wskazywać na występowanie obserwacji odstających

Wybór właściwego modelu analizy czynnikowej

- Wybór między PCA a FA zależy przede wszystkim od celu analizy
- W klasycznej analizie czynnikowej mała liczba czynników pozwala wyjaśniać zależności pomiędzy zmiennymi obserwowalnymi; chcemy zidentyfikować zmienne ukryte
- W analizie składowych głównych dążymy do zachowania jak największej ilości informacji przy jak najmniejszej liczbie nowych zmiennych; chcemy uprościć strukturę danych
- FA to analiza modelowa, PCA to technika eksploracyjna, pomocnicza